Une approche à base de proximité pour la détection de communautés egocentrées

Maximilien Danisch, Jean-Loup Guillaume, Bénédicte Le Grand

Une approche à base de proximité pour la détection de communautés egocentrées

Maximilien Danisch[1], Jean-Loup Guillaume[1] et Bénédicte Le Grand[2] †

[1]*LIP6, Université Pierre et Marie Curie, 4 Place Jussieu, 75005 Paris, France*
[2]*CRI, Université Paris 1 Panthéon-Sorbonne, 90 rue de Tolbiac, 75013 Paris, France*

Nous proposons ici une approche performante pour déplier la structure communautaire egocentrée sur un sommet d'un gaphe. Nous montrons que, bien que chaque sommet d'un réseau appartienne en général à plusieurs communautés, il est souvent possible d'identifier une communauté unique si l'on considère deux sommets bien choisis. La méthodologie que nous proposons repose sur cette notion de communauté multi-egocentrée ainsi que sur l'utilisation d'une mesure de proximité dérivée de techniques de dynamique d'opinion, la *carryover opinion*. Cette approche pallie les limites des fonctions de qualité traditionnellement utilisées pour la détection de communautés egocentrées, et consiste à étudier les irrégularités dans la décroissance de cette mesure de proximité. Cet article est une version courte de [DAN13].

Keywords: détection de communautés, communauté egocentrée, communauté multi-egocentrée, mesure de proximité

1 Introduction

De nombreux réseaux réels, tels que des réseaux sociaux ou des réseaux informatiques, peuvent être modélisés par des grands graphes, souvent appelés graphes de terrain. Ces réseaux réels ont été fortement étudiés ces dernières années, du fait de l'augmentation du nombre de jeux de données disponibles et du besoin de compréhension de tels systèmes dans de très nombreux contextes. La notion de communauté ‡ est centrale dans ce contexte et la recherche d'algorithmes efficaces pour identifier automatiquement de telles communautés constitue un défi, comme l'atteste l'article de synthèse de [FOR10].

Dans des systèmes réels, les communautés se chevauchent naturellement. Par exemple dans un réseau social chaque individu appartient à plusieurs communautés : famille, collègues, groupes d'amis, etc. La découverte de tous ces groupes dans de grands graphes est complexe et, outre le problème de l'efficacité du calcul, l'interprétation d'une structure en communautés recouvrantes peut être difficile. Certains travaux ont cependant abordé ce problème, par exemple [PAL05]. Une simplification consiste à se limiter à un partitionnement, dans lequel chaque sommet appartient à une et une seule communauté. Ce problème est également complexe et il n'y a pas encore de solution parfaite, mais de nombreuses définitions et algorithmes associés existent. La définition communément acceptée est la modularité [GIR02], qui privilégie les communautés plus denses qu'attendues et la méthode de Louvain [BLO08] fait partie des quelques méthodes très efficaces pour optimiser la modularité.

Une approche intermédiaire pour conserver le réalisme des communautés recouvrantes tout en simplifiant un peu le problème consiste à se concentrer sur un seul sommet et à tenter d'identifier l'ensemble des communautés auxquelles il appartient, aussi appelées communautés *egocentrées*. L'approche classique consiste à partir d'un groupe de sommets (qui ne contient initialement que le sommet considéré en général), puis à ajouter ou enlever des sommets à ce groupe itérativement pour optimiser une fonction de qualité donnée [CLA05, FRI11, NGO12]. Ce type d'approche souffre généralement de deux défauts majeurs : (i) il existe souvent des paramètres d'échelle cachés qui ont tendance à privilégier une certaine taille ou densité de communauté ; (ii) L'optimisation de la fonction de qualité est aussi très complexe du fait de la nature non-convexe de l'espace d'optimisation.

†. Ce travail est partiellement soutenu par l'Agence Nationale pour la Recherche, projet DynGraph ANR-10-JCJC-0202.
‡. Une communauté est intuitivement définie comme un groupe de sommets fortement connectés entre eux mais peu liés au reste.

Dans cet article nous proposons une approche orthogonale basée sur la détection d'irrégularités dans la décroissance d'une mesure de proximité entre sommets. Nous explicitons les résultats de notre approche en l'appliquant à un réseau réel contenant l'ensemble des pages de wikipedia (2 millions de pages) et les liens hypertextes entre ces pages (40 millions de liens) [PAL08].

2 Méthodologie

Notre approche nécessite l'utilisation d'une mesure de proximité pour évaluer la "proximité" entre deux sommets donnés et nous utilisons pour cela la *carryover opinion*, introduite dans [DAN12], pour sa rapidité d'exécution. Étant donné un sommet u, nous mesurons sa proximité à tous les autres sommets du graphe puis nous trions les proximités obtenues par ordre décroissant. Si l'on observe des irrégularités dans la décroissance de la courbe de proximité en fonction du rang cela met en évidence l'existence de communautés. Plus précisément, si plusieurs sommets sont également similaires au sommet u alors que le reste des sommets est peu similaire, on observe un plateau sur la courbe suivi d'une décroissance brusque, les sommets du plateau forment alors une communauté de u. Cependant, la proximité en fonction du rang décroit le plus souvent de manière régulière en suivant une loi sans échelle, i.e., aucune taille caractéristique ne peut être extraite de la mesure, ce qui signifie que le sommet appartient à plusieurs communautés de différentes tailles, ou n'appartient à aucune communauté.

Pour remédier à ce problème nous proposons de choisir un autre sommet v, et de chercher une communauté egocentrée à la fois sur u et sur v (la notion de communauté multi-egocentrée a été introduite dans [DAN12]). Pour cela nous calculons pour tous les sommets du graphe le minimum de leur proximité à u et à v, ce minimum représentant à quel point le sommet considéré est proche de u ET de v, comme illustré sur la figure 1. À nouveau, une irrégularité dans la décroissance des scores indique une communauté.

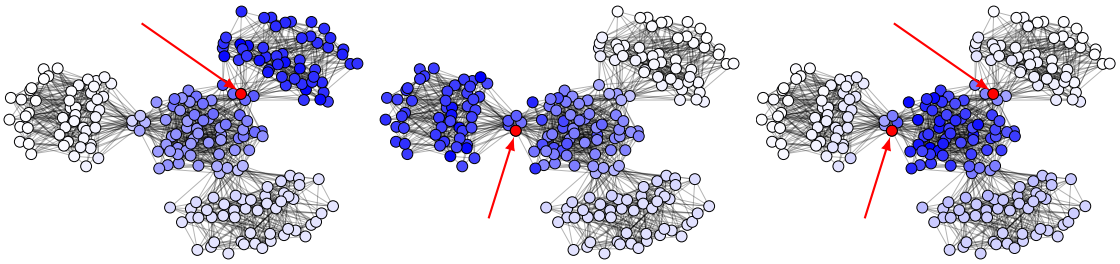


FIGURE 1: (1a) et (1b) : proximité de chaque sommet du graphe au sommet pointé par une flèche. Plus un sommet est sombre, plus son score est élevé. (1c) : minimum des scores de (1a) et (1b) qui permet de voir l'émergence de la communauté centrale. Le graphe considéré est l'union de 4 graphes aléatoires se chevauchant sur 5 sommets avec une probabilité d'existence de lien de $0,2$.

Plus en détail, pour un sommet u donné l'algorithme consiste à :

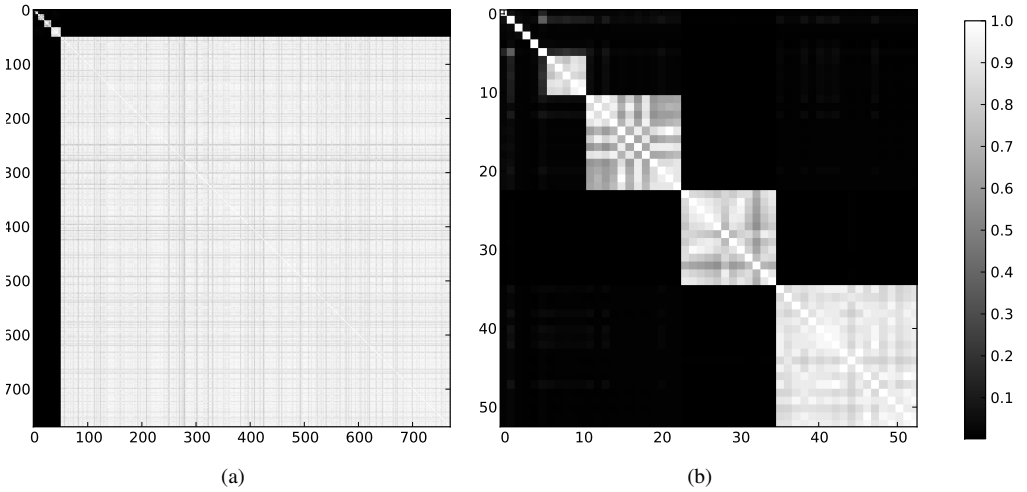
1. Choisir un ensemble de sommets candidats moyennement similaires à u. En effet si un sommet est trop peu similaire (resp. trop similaire) à u, il ne partagera aucune communauté (resp. beaucoup de communautés) avec u. Le but est que les deux sommets n'en partagent qu'une.
2. Pour chaque sommet candidat v, chercher une communauté multi-egocentrée sur u et v :
   - Pour chaque sommet du graphe, calculer le minimum de sa proximité à u et à v.
   - Trier les valeurs obtenues par ordre décroissant.
   - Si la pente maximale de la courbe de proximité en fonction du rang est plus grande qu'un certain seuil, alors une communauté est détectée, composée des sommets situés avant cette pente.
   - Si u appartient à la communauté détectée, alors la communauté est conservée, sinon elle est éliminée
3. Nettoyage et étiquetage des communautés trouvées :

– Plusieurs candidats peuvent donner naissance à des communautés très similaires. Afin d'éliminer ce bruit, on peut faire l'intersection des communautés très similaires afin de n'en garder qu'une.
– À l'inverse, si une communauté n'est similaire à aucune autre elle est éliminée. Cela se produit en effet quand la pente maximale détectée est faible et on peut donc assimiler cela à une erreur.
– La communauté reçoit le label du sommet le mieux classé.

3 Résultats et validation

Nous présentons ici les résultats pour un unique sommet du réseau wikipedia, la page *Chess Boxing* [§] pour laquelle les résultats sont très pertinents, facilement interprétables et validables manuellement.

Pour le sommet "Chess Boxing", nous avons utilisé l'algorithme présenté précédemment sur 3000 sommets candidats choisis aléatoirement entre le centième et le dix-millième sommet parmi les plus similaires à "Chess Boxing". Sur les 3000 tentatives, 770 ont abouti à l'identification de plateaux suivis de décroissances nettes, fournissant autant de communautés. La figure 2 présente la matrice des similarités de Jaccard (similarité entre deux ensembles) entre ces 770 groupes. Sur ces 770 communautés, la phase de nettoyage élimine directement 6 communautés qui ne sont similaires à aucune autre et sont donc assimilées à des erreur de l'algorithme de détection de pente (une vérification manuelle confirme que ces communautés n'ont pas de sens).



(a) (b)

FIGURE 2: (2a) : matrice des similarités de Jaccard des 770 communautés, réordonnée pour que les groupes similaires soient proches. Un groupe de 716 communautés émerge. (2b) : zoom sur la partie supérieure gauche de la matrice permettant de mieux voir les autres groupes de communautés. On y observe 4 autres groupes de communautés et 6 communautés très peu similaires aux autres.

On détecte également 5 groupes de communautés très similaires, mais peu similaires aux autres. L'intersection des communautés à l'intérieur de chaque groupe donne une communauté étiquetée (avec le label du sommet le mieux classé). Les 5 communautés identifiées sont : "Enki Bilal" (35 sommets), "Uuno Turhapuro" (26 sommets), "Da Mystery of Chessboxin' (254 sommets)", "Gloria" (55 sommets) et "Queen's Gambit" (1619 sommets). Toutes ces communautés sont clairement liées au Chess Boxing tout en étant sur des sujets très différents.

De manière surprenante l'algorithme ne trouve aucune communauté liée à la boxe. Ceci pourrait être un problème de l'algorithme, mais la page wikipedia de "Chess Boxing" indique que la plupart des pratiquant ont un passé dans les échecs et apprennent la boxe par la suite. Ils peuvent donc être plus importants dans le monde des échecs que dans celui de la boxe. Ceci pourrait expliquer que le sommet "Chess Boxing" soit dans la communauté des échecs, mais à la limite de celle de la boxe.

§. Le ChessBoxing est un sport mêlant échecs et boxe avec des rounds alternés.

4 Conclusion et perspectives

Nous avons proposé un algorithme qui permet d'identifier et d'étiqueter les communautés egocentrées pour un sommet d'un graphe. Notre approche est basée sur la recherche d'irrégularités dans la décroissance des valeurs d'une mesure de proximité. L'algorithme est efficace en temps et permet de trouver les communautés d'un sommet sur des graphes contenant des millions de sommets. En utilisant la notion de communautés multi-egocentrées, l'algorithme identifie dans un premier temps des sommets candidats pouvant permettre l'identification de communautés, puis cherche des communautés centrées sur notre cible et sur ces candidats, et procède enfin à une phase de nettoyage et d'étiquetage des communautés.

Bien que cet algorithme soit performant en l'état, de nombreuses pistes d'amélioration sont possibles. Tout d'abord, la détection de communautés se base sur la recherche de plateaux et de décroissances fortes. La méthode actuelle peut être améliorée, notamment par la recherche de plusieurs décroissances, ce qui permettrait de trouver plusieurs communautés à des échelles différentes pour un même candidat.

De plus, l'algorithme utilise pour l'instant une notion de communauté bi-centrée or, il est possible que certaines communautés n'apparaissent que centrées sur 3 sommets ou plus. Cette généralisation doit être validée sur des exemples de petites tailles car le temps de calcul sera fortement augmenté à moins d'améliorer très significativement la méthode de sélection des candidats. Une approche pourrait être de considérer que si un candidat v a fourni de bons résultats alors des sommets qui lui sont très similaires n'apporteront pas de nouvelle information.

Cette notion de rapidité de l'algorithme est centrale pour pouvoir suivre l'évolution des communautés sur plusieurs instants il est important que les calculs soient aussi efficaces que possible.

Enfin, nous avons observé que l'algorithme peut avoir des difficultés à identifier de petites communautés si elles sont proches de grosses communautés. Pour cette raison, tenter d'appliquer l'algorithme à des sommets très populaires tels que "Biology" ou "Europe" ne conduit qu'à une grosse communauté, alors que l'on s'attendrait à trouver divers sous-domaines de la biologie ou différents pays européens. Une piste d'amélioration pourrait consister à relancer récursivement l'algorithme sur des communautés identifiées pour trouver des sous-communautés.

Références

[DAN13] M. Danisch, J.-L. Guillaume and B. Le Grand. Unfolding Ego-Centered Community Structures with "A Similarity Approach". Complex Networks IV, 2013, pages 145–153, Springer.

[DAN12] M. Danisch, J.-L. Guillaume and B. Le Grand. Towards multi-ego-centered communities : a node similarity approach. Int. J. of Web Based Communities (2012)

[BLO08] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre. 'Fast unfolding of communities in large networks'. J. Stat. Mech. (2008).

[CLA05] Aaron Clauset. 'Finding local community structure in networks'. PHYSICAL REVIEW E 72, 026132, 2005.

[FOR10] Santo Fortunato. Community detection in graphs. Physics Reports 486, 75-174 (2010)

[FRI11] Adrien Friggeri, Guillaume Chelius, Eric Fleury. 'Triangles to Capture Social Cohesion'. IEEE (2011).

[GIR02] M. Girvan and M. E. J. Newman. 'Community structure in social and biological networks'. PNAS June 11, 2002, *Biometrika*, vol. 99 no. 12, pp. 7821-7826.

[NGO12] Blaise Ngonmang, Maurice Tchuente, and Emmanuel Viennet. 'Local communities identification in social networks'. Parallel Processing Letters, 22(1), March 2012.

[PAL05] Palla, G., I. Derenyi, I. Farkas and T. Vicsek. 'Uncovering the overlapping community structure of complex networks in nature and society'. Nature 2005.

[PAL08] Gergely Palla, Illes J. Farkas1, Peter Pollner, Imre Derenyi and Tamas Vicsek. 'Fundamental statistical features and self-similar properties of tagged networks'. New J. Phys. 10 123026 (2008)